Preprint
UCRL-JC-145659

A Proposal for Establishing a Free Market Basis for Plant Genome Information Exchange

T. Slezak

This article was submitted to
National Science Foundation Plant Genome Initiative Workshop,
Rockville, MD, September 11, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

September 26, 2001

DISCLAIMER

This document was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor the University of California nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or the University of California. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or the University of California, and shall not be used for advertising or product endorsement purposes.

This is a preprint of a paper intended for publication in a journal or proceedings. Since changes may be made before publication, this preprint is made available with the understanding that it will not be cited or reproduced without the permission of the author.

This report has been reproduced directly from the best available copy.

Available electronically at http://www.doe.gov/bridge

Available for a processing fee to U.S. Department of Energy
and its contractors in paper from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831-0062
Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-mail: reports@adonis.osti.gov

Available for the sale to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/ordering.htm

OR

Lawrence Livermore National Laboratory
Technical Information Department's Digital Library
http://www.llnl.gov/tid/Library.html

NSF Plant Genome Informatics Workshop White Paper

Tom Slezak, LLNL, slezak@llnl.gov, 925-422-5746

A Proposal for Establishing a Free Market basis for Plant Genome Information Exchange

Abstract

The current situation of genomics information exchange is reminiscent of some Third World cities, where residents despair they will ever get official utility service and therefore tap into power, phone, and gas sources with makeshift connections. Thus, each genomics grant spawns yet another idiosyncratic Web site, with makeshift links to whatever random Web sites the PI is most familiar with. There are few standards for semantics of data, and fewer standards for automating the interchange or integration of these autonomous Web sites. The US Plant Genome Initiative (PGI) has been an enthusiastic contributor to this proliferation of chaotic Web sites, but to its credit it appears to be the first major program to attempt to find a solution.

Some of us from the earliest days of the Human Genome Program have been acutely aware of the problems of genomic data integration, since long before the Web appeared and made the problem exponentially harder to resolve. We have seen large scale attempts, and subsequent failures or inadequacies, of many potential solution approaches (i.e., database federation, classical data warehousing, centralized data, etc.) and believe we know at least some of the reasons they still remain inadequate.

It is our opinion that the only solution that has a chance of succeeding is one that considers the overall economics of genomics data production, sharing, and integration. We believe that attempting to create a kind of Free Market for data created under the Plant Genome Initiative will represent the most practical, powerful, and cost-effective approach to dealing with the broad range of plant genome information that has been unleashed.

Central to creating this Free Market is the creation of minimal standards required for access to any Web-enabled data site funded by the PGI.; a set of URLs that can be accessed via either interactive Web browsers or via programs running on other Web sites. Proper design and use of the standard URLs would aid in automating proper attribution for information extracted from other sites. The availability of this *lingua franca* for information access will permit the creation of an information market where grants can be funded for hierarchies of data integration and/or data mining that otherwise would not be feasible to even propose. We will propose such a set of URLs below.

Our many years of experience in this field make us painfully aware that there are no silver bullets for this most challenging and important of information problems. We are not claiming that a Free Market approach to Plant Genome information will be painless or without major problems, but considering the sociology and economics of the

Genomics domain, we think that it is the only approach that has even a chance of succeeding.

Background

It is, at least in part, the immature state of genomics in general that led to the current problematic situation outlined above. Basic terminology is still capable of causing controversies (especially when you are trying to build a database) and the most fundamental of ontologies are still in early states of development. Almost unbelievably, Genbank remains a flat file with amorphous text fields for much of the annotation that could and should be stored in a searchable relational format. There are no easy ways to query seamlessly across even the 3 or 4 most commonly-used Web data resource, much less across any *ad hoc* collection of Web sites.

Before we present our Proposal on how to wring order out of this chaos, we need to briefly discuss some other approaches that have been tried and (in our opinion) have proven to be inadequate for various reasons.

- **Language-based approaches**

A very "computer science" approach pursued since the early 1990s was to define a language capable of representing arbitrarily complex genomics queries that crossed multiple data sources. The Kleisli and K2 languages from the University of Pennsylvania are examples of this approach. Despite extensive publicity in the bioinformatics community for many years, this technique never achieved popular status. However, some genomics companies are still using it.

- **Flat file/text retrieval/search engine systems**

Historically, many legacy genomics resources started off as flat-file systems, due to the real and perceived difficulties of using SQL and a general reluctance to spend money on frills like real databases. As a result, a number of data retrieval systems have sprung up based on searching/linking flat files. The SRS system is an example of such a system that has become popular in Europe. People have also used techniques from text retrieval and search engine domains. All of these techniques have things to offer in limited situations, but fail in general due to semantic differences, the inherent difficulties of parsing unstructured data, and the fact that keywords are inadequate to represent all the relationships between genomics sources.

- **Database federation, distributed databases**

Several serious attempts at database federation and distributed databases were mounted in the past decade. The EU's Integrated Genome Database (IGD) came the closest to success. Led by Otto Ritter, it featured a federation-wide normalization of semantics and objects, thus avoiding one major reason for integration failure. At its zenith, IGD integrated about a dozen federation databases. It eventually collapsed due to scaling problems related to maintaining the interfaces needed to exchange data among all the varied resources. On average, each genomic database goes through about 2 new schema releases per year. According to Otto, this meant that, on average, IGD broke every two weeks.

A similar attempt at federating the DOE-funded databases was initiated, but never really achieved critical mass. This was mostly due to philosophical differences between autonomous sites that were never brought to resolution. The immaturity of the domain was also a contributing factor.

- **Classical data warehousing**

In recent years there have been some attempts at using data warehousing techniques to tackle genomic data integration. The author was involved with the DataFoundry effort at LLNL, which brought Swiss Prot, PDB, Scop, Cath, and dbEST together into a unified warehouse schema. Although a novel use of meta-data largely automated the procedures of generating wrappers and parsers for external data sources, this approach ultimately also proved unscalable for the explosion of relevant Web-enabled genomic data resources.

- **Centralization**

Another tempting solution has always been to centralize all relevant data into one place; what some might call the "NCBI approach" to genomics data. Given the current explosion of genomic data, centralization is probably not an option for the Plant Genome Initiative; there is simply too much important data out there for anybody to centrally manage and control. Another flaw of centralization is the tendency to build interfaces that only permit querying from the points of view that the developer considered important (e.g., the inability of *Entrez* to query annotation as a first-class object).

- **Web robots/agents**

Given that the Web has in some sense caused the current problem, or at least made it far worse, it is not surprising that we must look to some sort of Web-based method for its solution. One approach that is probably not suitable for the PGI is the kind of Web robot or intelligent agent technology that has grown up to support Internet shopping and auction activities. The basic idea of these tools is that they are programs that "crawl" the Web looking for sites that support the target activity (say, shopping for children's toys.) They parse the HTML of sites to see if it appears that the site is in the proper domain, and if it has an interface that allows one to search the site. They repeatedly attempt to drill down to discover a form that will allow one to enter the name of the desired object, and hopefully elicit a price. The key point to note here is that the target sites are presumed to be totally passive to the robot and are not providing any special information to aid it. One can imagine such a robot searching for BLAST sites on the Web, for links to gene names, for clone numbers, or for cDNA clone names. Where these technologies fall short is that integrating genomic information is not simply analogous to Internet shopping. Techniques suitable for finding the cheapest Barney doll can only go so far towards solving genomic queries that need to span multiple domains and data resources.

Perhaps not surprisingly, this brief examination of several categories of techniques for data integration yields the result that that is no single Panacea to solve the problems we all face in the genomics domain.

Summary of our Strawman Proposal

With the above background in mind, here are components of a strawman proposal for how the PGI might approach solving their problems of getting the most bang for their bioinformatics buck:

1. **common data release policy**

   All PGI-funded projects must adhere to the same rules for data release. Sooner is better than later, and more is better than less. Sequencing sites must be held to the same rules for timing and the quality/quantity of releases. The PGI must decide how early draft contigs should be released, how soon "completed" sequences should be released (before or after "annotation" completes?), and whether a site can withhold sequence of completed work indefinitely awaiting publication. A more subtle issue here is that the public release of data funded by the PGI is more important than the appearance of a journal paper; a site's release history should be as important as its publication history when considering awards or renewals.

2. **cooperation of PGI sites mandated by funding agencies**

   All PGI-funded projects must be required to adhere to all aspects of the data sharing plan. It is up to the funding agencies to ensure that grantees have requested sufficient support to meet the informatics obligations.

3. **common ontologies to be developed for basic objects, attributes, annotation**

   The PGI-funded community must speak a common language for any meaningful data integration to be possible. This means that terms like "contig", "gene", and "finished sequnece" need to be defined and available to all grantees. It also means that community efforts such as the Gene Ontology (GO) project should be joined and enhanced for PGI purposes. This is perhaps the most difficult step of the entire plan, and should be aggressively fast-tracked by the PGI.

4. **common data interchange methodologies**

   Technology advances of the past decade have made irrelevant the details of how an information-generating site stores their data. It is also pretty obvious now that techniques like the usage of XML are the most sensible way to exploit the ontologies of (3) to be able to exchange queries and receive results. The PGI needs to have members become active in the relevant XML standards and development arenas.

5. common set of URLs implemented at each site to allow automated query access

The author proposed this idea in a talk given to the International Rice meeting in February, 2000. Shortly thereafter, he found that Lincoln Stein had proposed a much-improved variant in his Distributed Annotation System (DAS, http://stein.cshl.org/DAS) which solves the specific problem of having multiple independent sets of annotation upon a common reference sequence.

For the PGI, a combination of both ideas is required. A limited set of URLs would need to be implemented at each PGI-funded site, to allow the retrieval (4) of standard objects (3) of all PGI-funded data (2) in timely fashion (1). All URLs should be capable of being driven by a human at a browser, or by a program. Results in either HTML or XML should be a parameter option on all the standard URLs.

Here are a few examples of the basic URLs that need to be supported to make this work. These are not intended to be complete, only to stimulate discussion:

- Define the domain(s) of the site, in terms of Organisms (rice, wheat, grasses, ...), types of data contained (Sequence, Expression, Mapping, Comparative,...) or other appropriate dimensions.
- Define the standard objects (item 3 above) "keys" that can be accepted as query selection specifications, using a simple boolean query language supported by PGI URLs (example: return sequence, annotation where organism=rice and gene_name=XYZ). The concept here is to enforce an ability to query that goes beyond the very limited abilities of Genbank, without getting bogged down in an overly complex query language or enforcing any particular underlying mechanism for implementation. Not all sites will support all possible query complexities or objects..
- Define the standard objects (item 3 above) "results" (attributes, etc.) that can be returned to the requesting site when presented with one or more of the query "keys" specified. Note that not all sites will support all of the details from the ontologies above for queries or results.
- The standard query URL itself that accepts a boolean combination of query "keys" and returns XML or HTML formatted "results".

6. only define the basic data access and interchange; let the market provide all forms of higher integration

Requiring each PGI-funded site to adhere to these standards effectively establishes a "free market" of genomics data that can be exploited by others who have novel ideas for analyzing and presenting data integrated across multiple PGI sites. It also allows information providers equal access to do their own integration and comparison.

7. provide grants for information integration services and tools

It should be reasonably obvious by now that we are rapidly moving towards a time when a large portion of biology funding will be consumed by bioinformatics efforts. The PGI needs to realize that careful and creative funding of information integration tools and services are at least as important as the production of raw sequence and expression data. Creating a free market for information isn't enough by itself; a vigorous effort must be made to reward successes and prune failures in bioinformatics. As a trivial example, the PGI might wish to fund a central archival site, that extracted and maintained a copy of all other funded participant sites, to insure against loss or destruction of a member site, and to provide better performance for automated queries against PGI data. Note that it would not be necessary for this archival site to perform any active information integration, although doing so would presumably add value to the archiving function.

8. become active in genomics standards and integration efforts

Adopting a system like the proposed strawman would put the PGI out in the forefront of the bioinformatics community in terms of a coordinated plan of information release and free access. The PGI would need to be responsible to the rest of the genomics world and become an aggressive participant in getting other aspects of the global community to adopt and improve these standards.

Discussion

It is our contention that the only solution to the growing problems of isolated genomics data resources is to establish a free and level access to the data. We present a concept for doing this via browser or programmatic access, along with constraints for holding the entire PGI-funded community to it. One can readily imagine extending this concept to other communities by denying access to PGI sites unless their site also plays by the same rules, and by the establishment of "good neighbor" rules to avoid denial of service issues. (For example, "No more than K small or N large queries per second/hour".) It also should be clear that the set of standard URL queries will probably only return a relatively small fraction of the summary data from any active genomics production site. (For example, returning the sequence of a contig clearly should be possible via this system, whereas it is dubious at best if it should also be possible to extract all the raw data that went into that contig.)

Note that the main point of this concept is not how to *solve* the problems of doing information integration, but to establish a framework that makes it *possible* to achieve a solution. As shown here, the technology to establish such a framework is understood and can easily be implemented. I will leave it to others in this workshop to discuss the many interesting technical options for actual information integration.

Without a community-wide enforced standard of information release timeliness, semantics, and exchange format it will be futile to expend resources for massive information integration. Grantees need to make a mental shift to think of themselves as being part of a much larger process of knowledge gathering and information than just the small portion for which they are currently funded. Under the approach we are suggesting, grants would fall three categories: 1) producing sequence data, 2) developing information integration tools 3) conducting research utilizing the data and tools. This division of labor will discourage good biologists from becoming poor bioinformaticists and will encourage the development of a heirarchy of information integrators, analyzers and visualizers.

The current PGI granting process virtually ensures that half-hearted, under-funded informatics exists for every biology grant. The fact that this workshop is being held shows that it is time for the funding agencies to take a bold step to separate the raw presentation of grant output on the Web from the much harder task of intelligently integrating it across the span of genomics resources.

The author wishes to thank Nisha Mulakken and Bert Weinstein for their careful review of an early draft and their many constructive suggestions that I have incorporated. The NSF is also thanked for having the fortitude to tackle the problems of data integration.

This work was performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48.